Major Challenges to Achieve Exascale Performance

Shekhar Borkar
Intel Corp.
April 29, 2009

Acknowledgment: Exascale WG sponsored by Dr. Bill Harrod, DARPA (IPTO)

Outline

Exascale performance goals

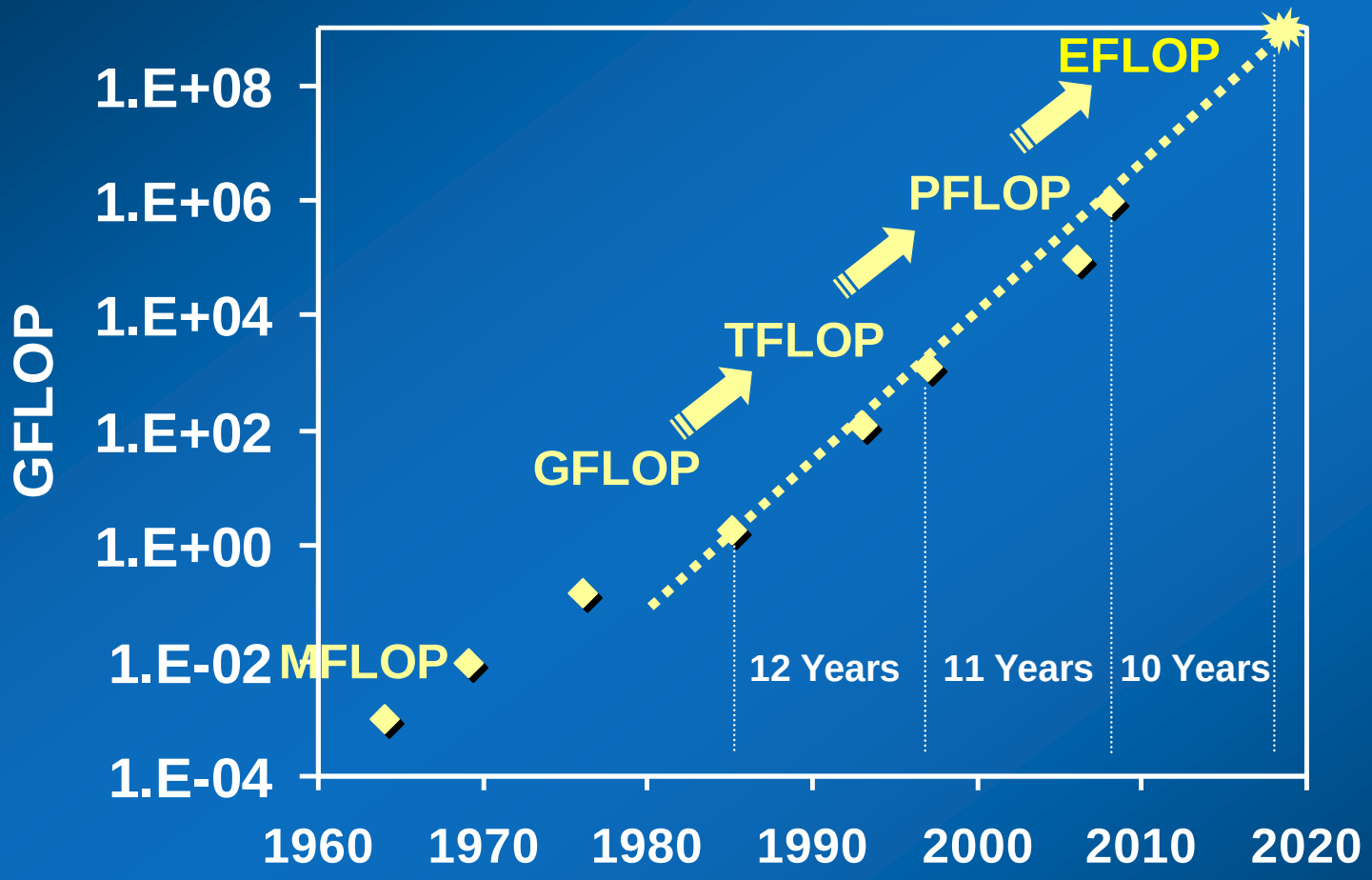
Major challenges

Potential solutions

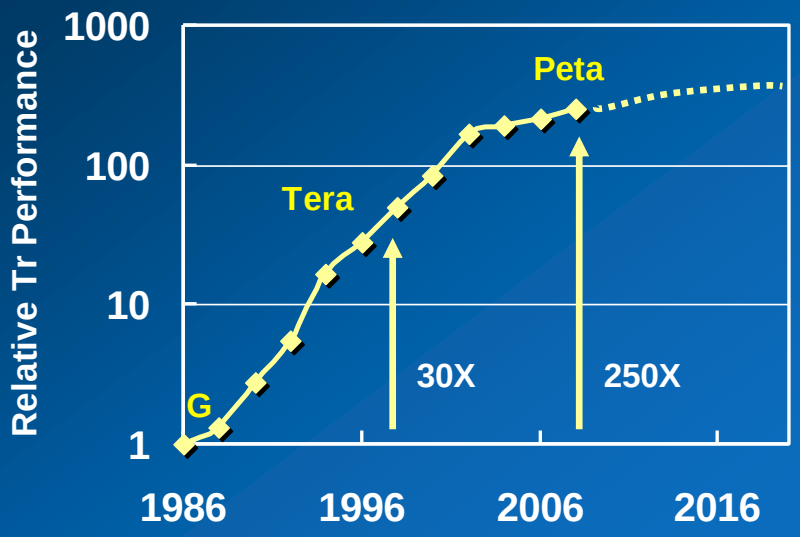
Paradigm shift

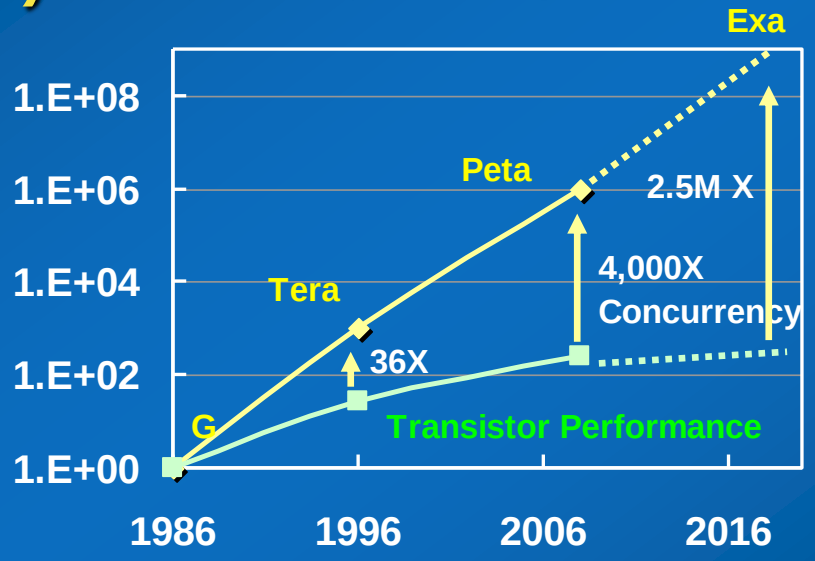
Summary

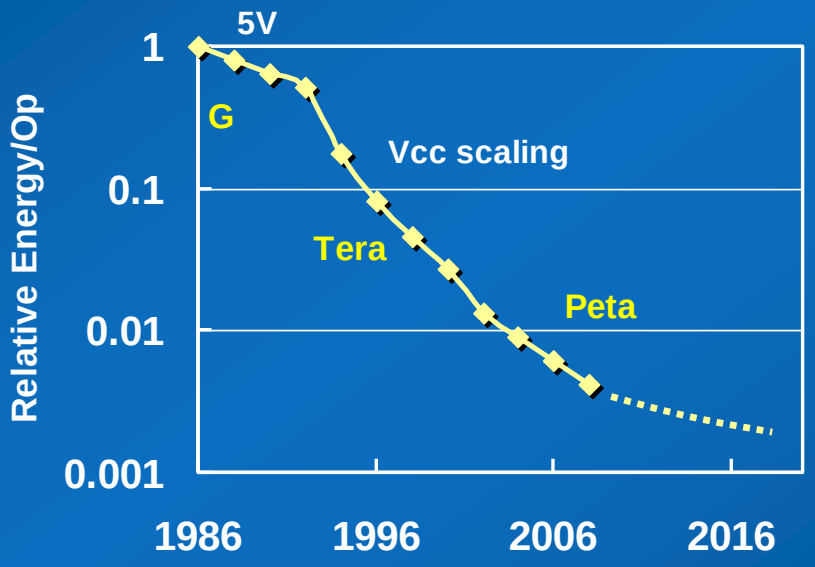
Performance Roadmap

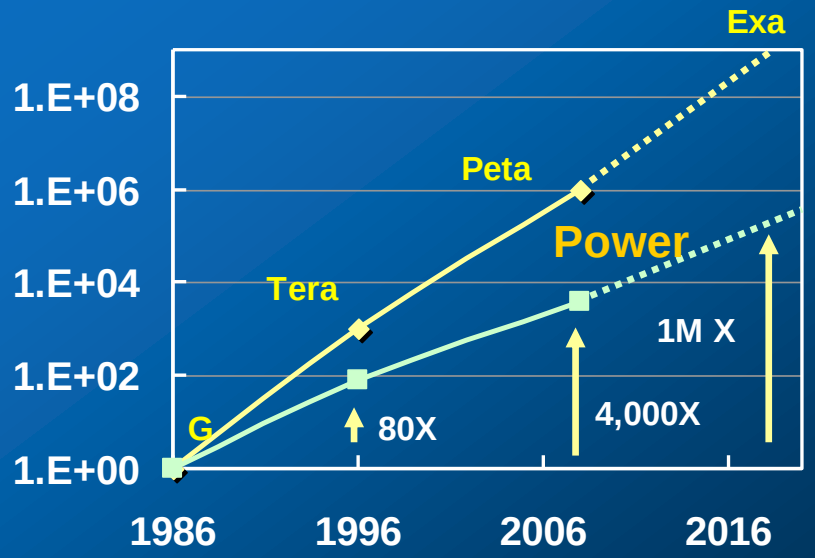


From Giga to Exa, via Tera & Peta



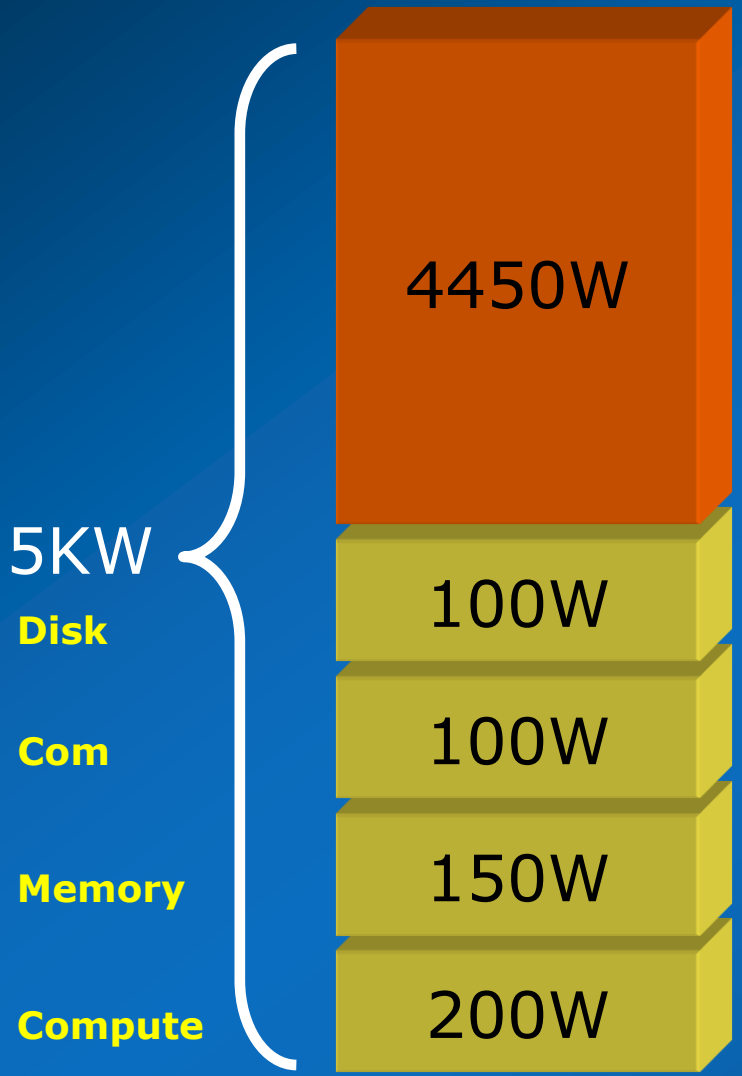






Building with Today's Technology

TFLOP Machine today



Decode and control
Translations
…etc
Power supply losses
Cooling…etc

10TB disk @ 1TB/disk @10W

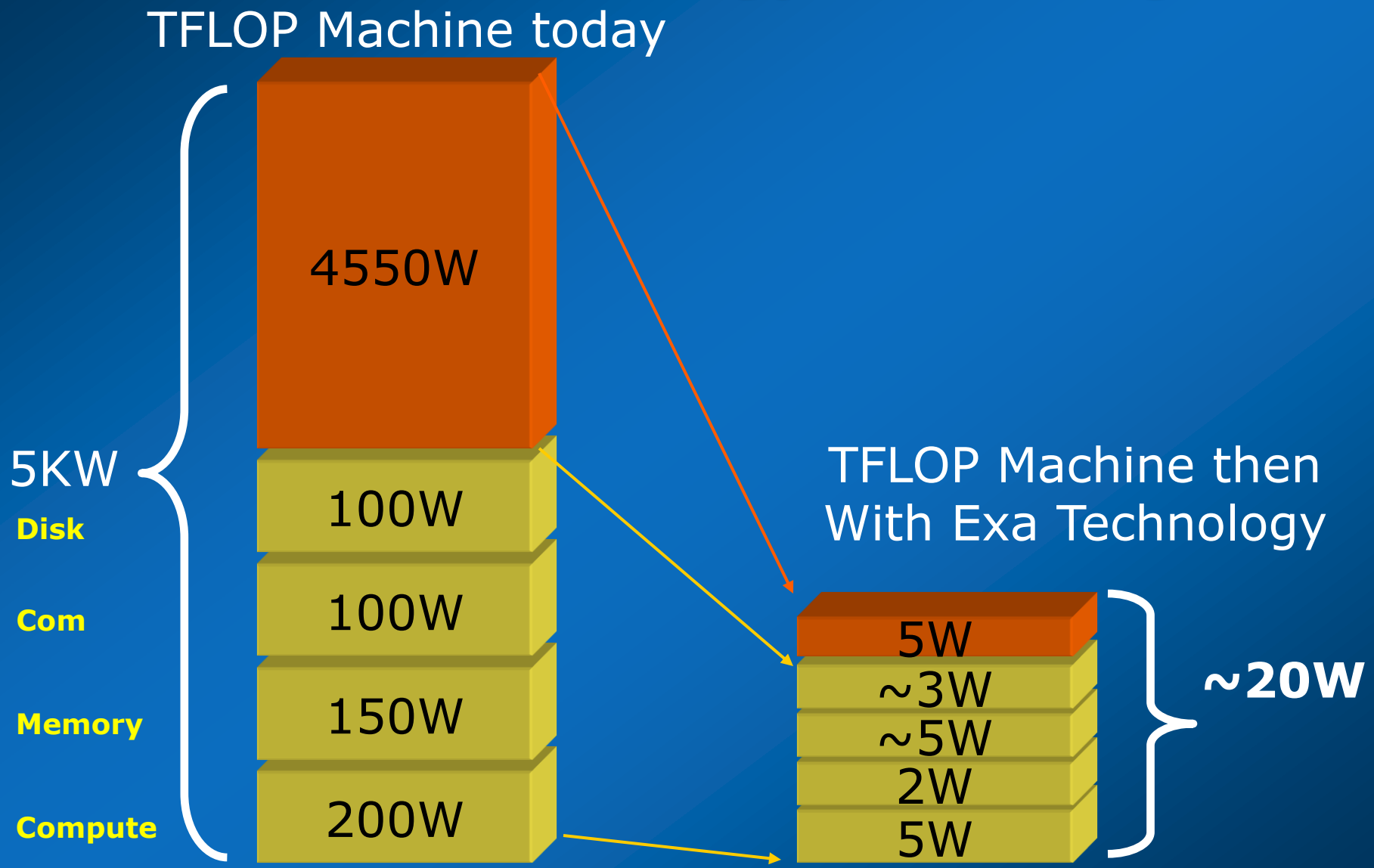
100pJ com per FLOP

0.1B/FLOP @ 1.5nJ per Byte

200pJ per FLOP

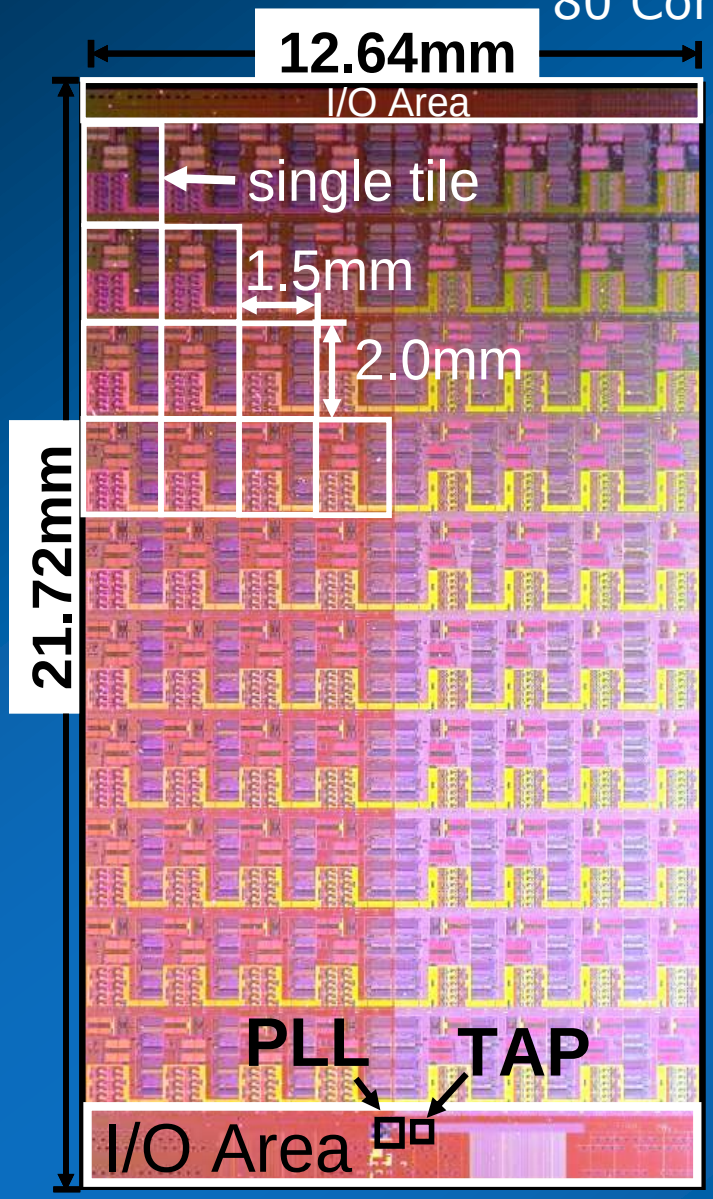
KW Tera, MW Peta, GW Exa?

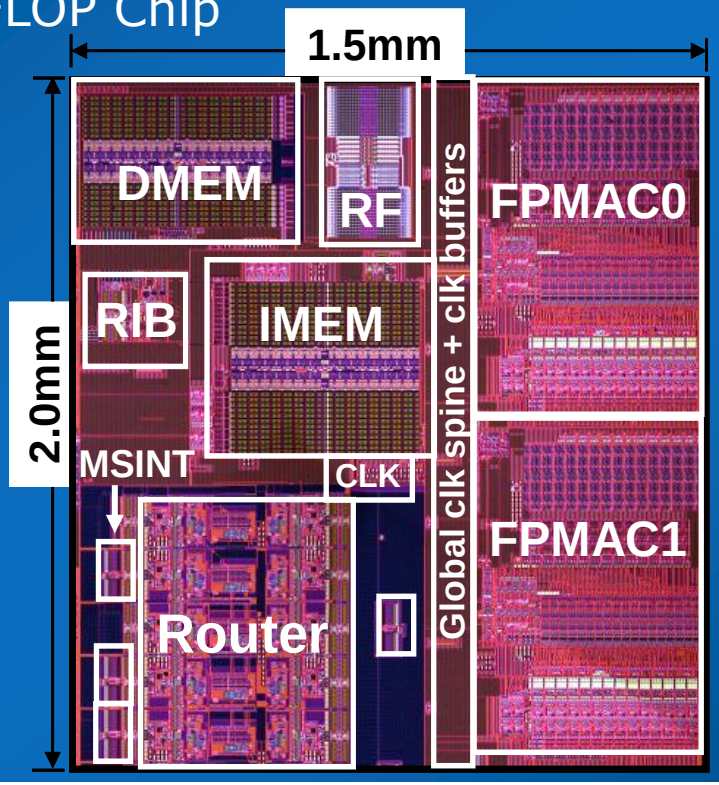
The Power & Energy Challenge



Starting Point: Optimistic yet Realistic

80 Core TFLOP Chip



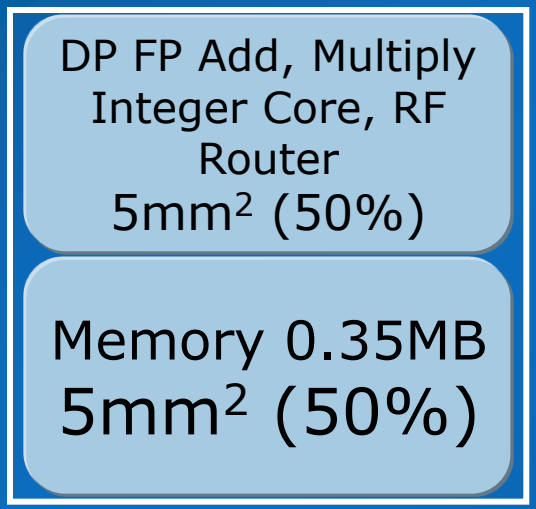


Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm^2
Tile area	3mm^2
Package	1248 pin LGA, 14 layers, 343 signal pins

Scaling Assumptions

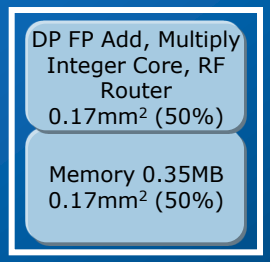
Technology (High Volume)	45nm (2008)	32nm (2010)	22nm (2012)	16nm (2014)	11nm (2016)	8nm (2018)	5nm (2020)
Transistor density	1.75	1.75	1.75	1.75	1.75	1.75	1.75
Frequency scaling	15%	10%	8%	5%	4%	3%	2%
Vdd scaling	-10%	-7.5%	-5%	-2.5%	-1.5%	-1%	-0.5%
Dimension & Capacitance	0.75	0.75	0.75	0.75	0.75	0.75	0.75
SD Leakage scaling/micron	1X Optimistic to 1.43X Pessimistic						

65nm Core + Local Memory



10mm2, 3GHz, 6GF, 1.8W

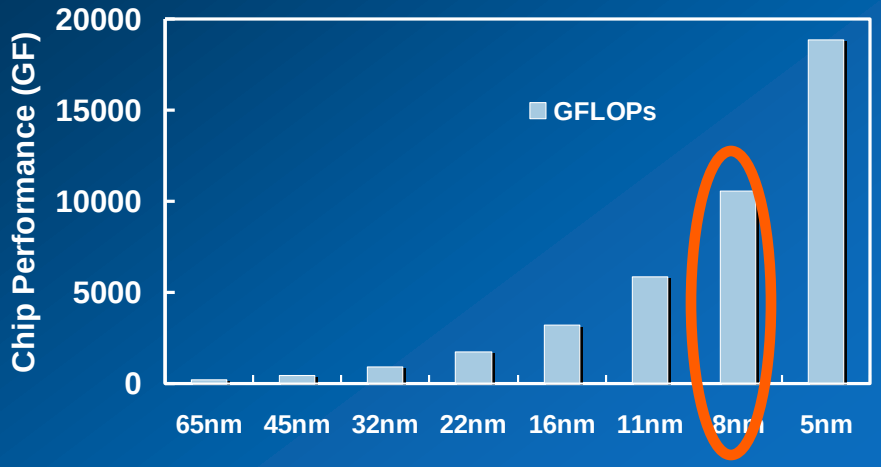
8nm Core + Local Memory

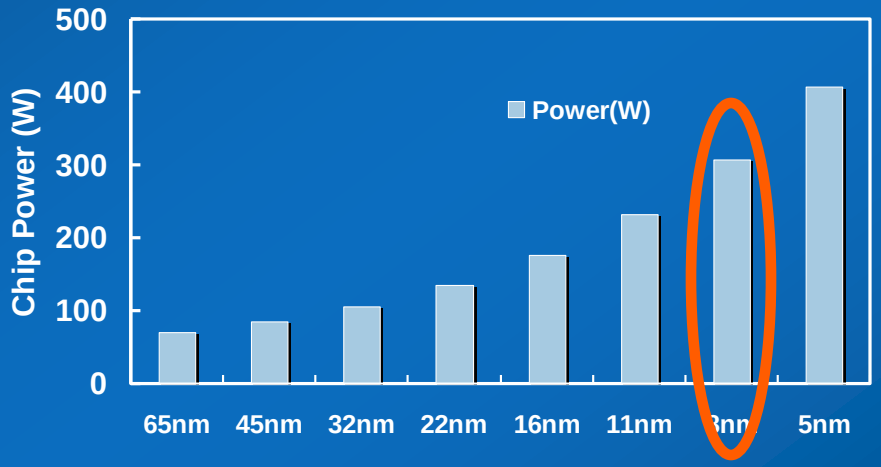


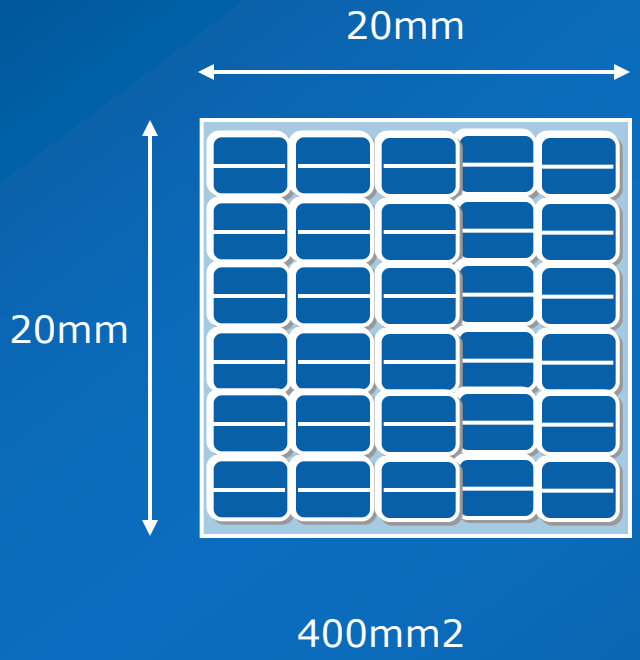
~0.6mm

0.34mm2, 4.6GHz, 9.2GF, 0.24 to 0.46W

Processor Chip





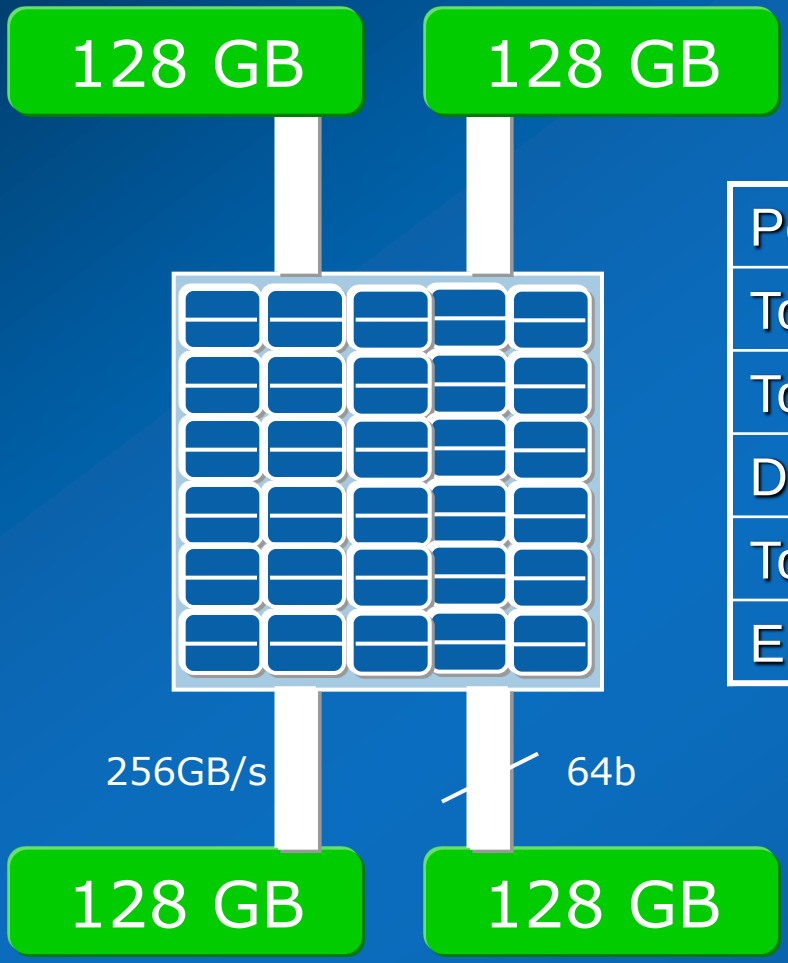


2018, 8nm technology node

Cores/Module	1150
Total Local Memory	400 MB
Frequency	4.61 GHz
Peak performance	10.6 TF
Power	300 - 600W
Energy efficiency	34 - 18 GF/Watt

30-60 MW for Exascale

Processor Node



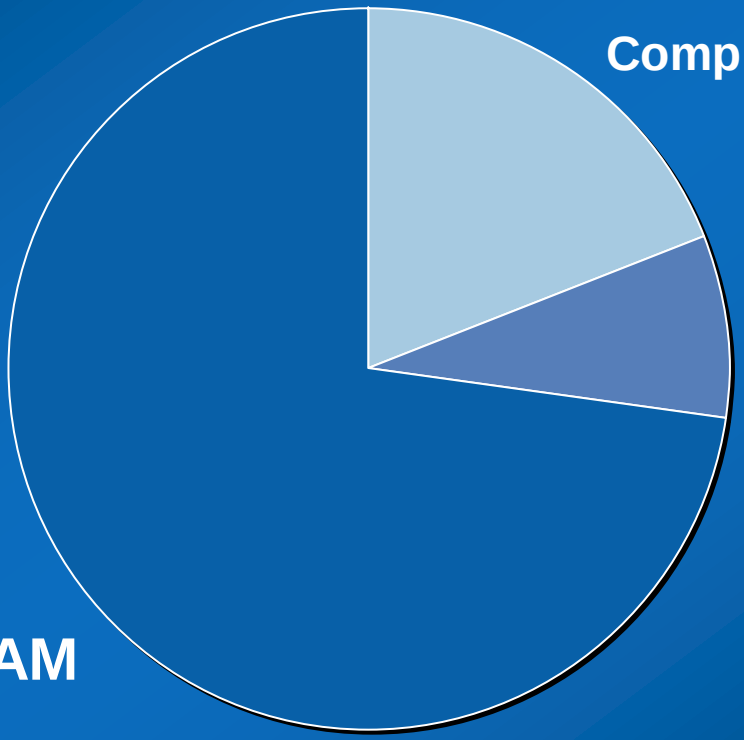
Peak performance	10.6 TF
Total DRAM Capacity	512GB
Total DRAM BW	1TB/s (0.1B/FLOP)
DRAM Power	800 W*
Total Power	1100 - 1400W
Energy efficiency	9.5 - 8 GF/Watt

110-140 MW for Exascale

*Assumes 5% Vdd scaling each technology generation
140 pJ energy consumed per accessed bit

Node Power Breakdown

10 TF, ~ 1KW



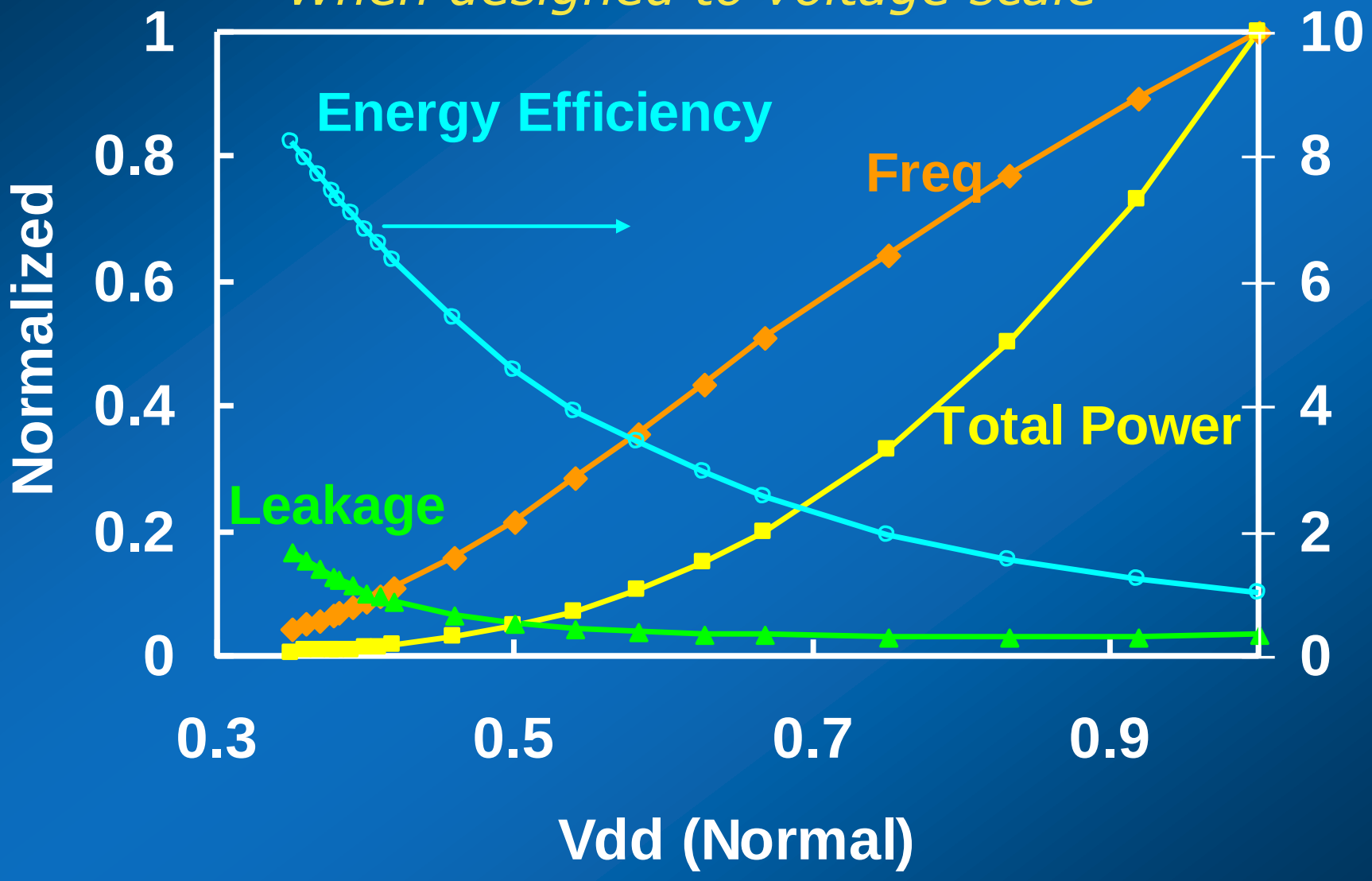
Aggressive voltage scaling

Hierarchical heterogeneous topologies

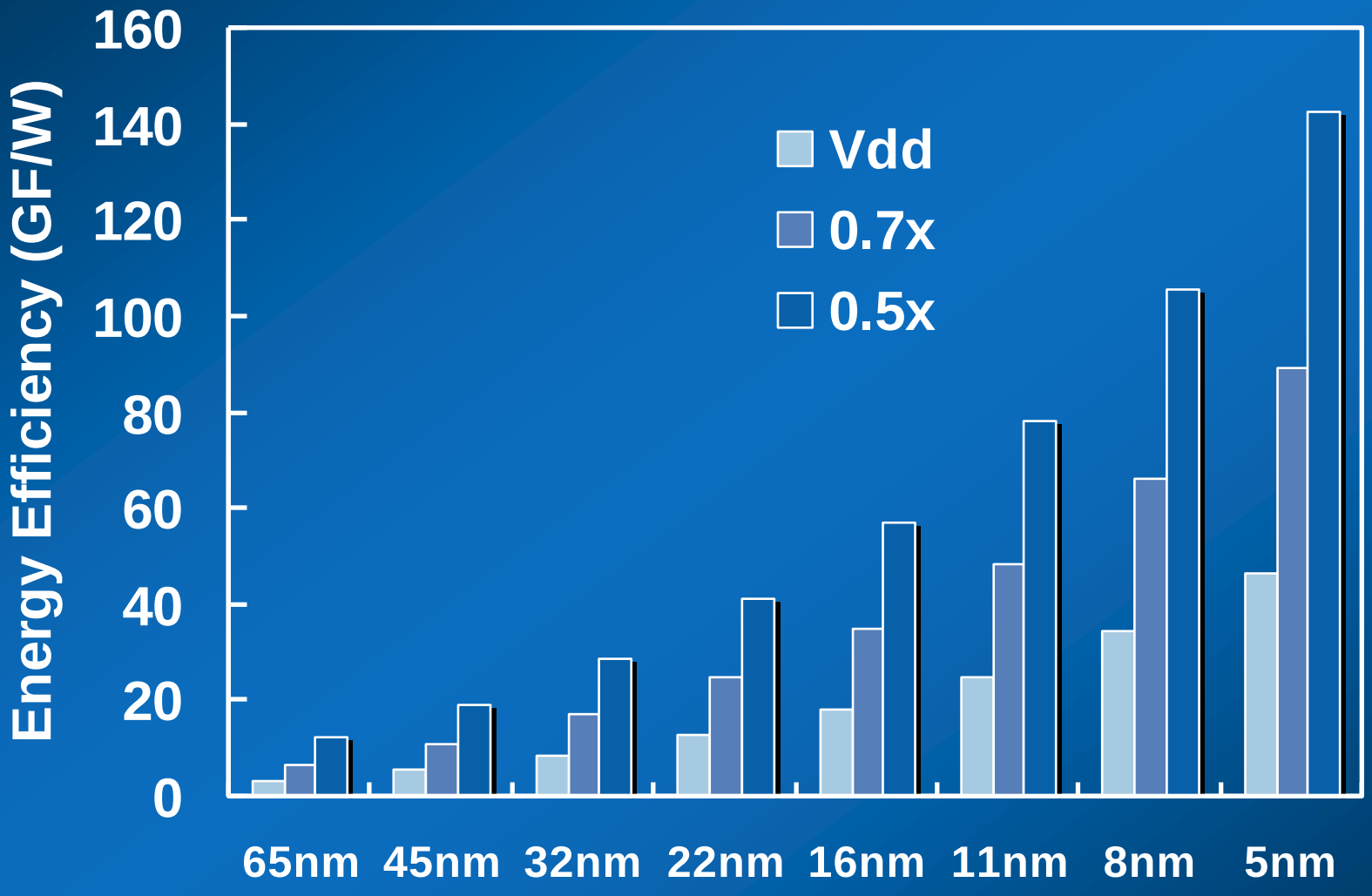
Efficient signaling Repartitioning

Voltage Scaling

When designed to voltage scale

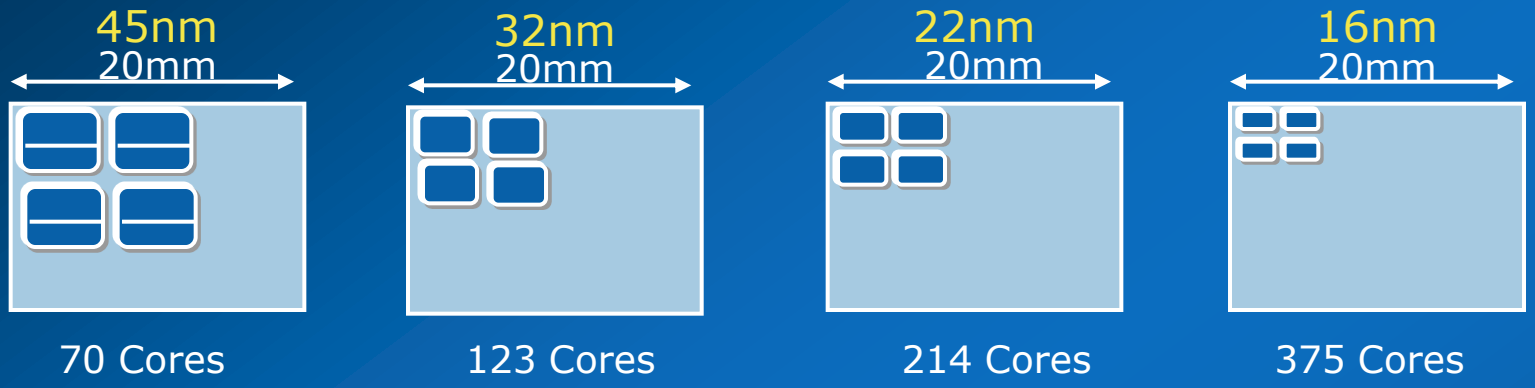


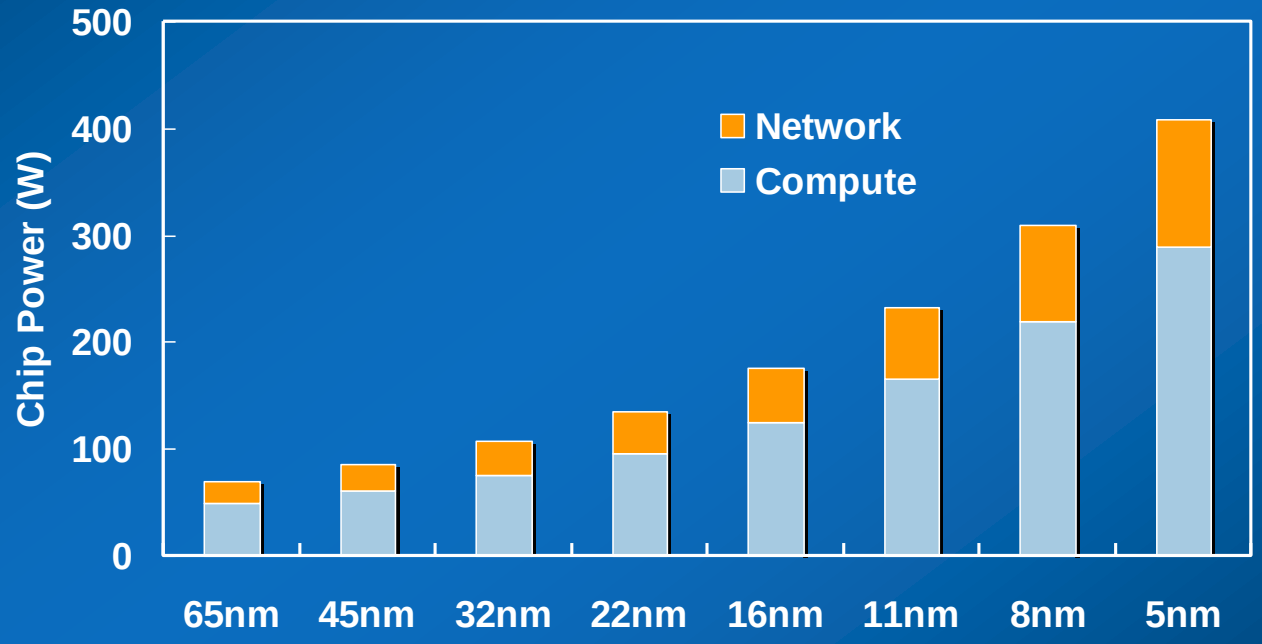
Energy Efficiency with Vdd Scaling



~3X Compute energy efficiency with Vdd Scaling

On-die Mesh Interconnect





On-die network (mesh) power is high
Worse if link width scales up each generation

Mesh—Retrospective

Bus: Good at board level, does not extend well

- Transmission line issues: loss and signal integrity, limited frequency
- Width is limited by pins and board area
- Broadcast, simple to implement

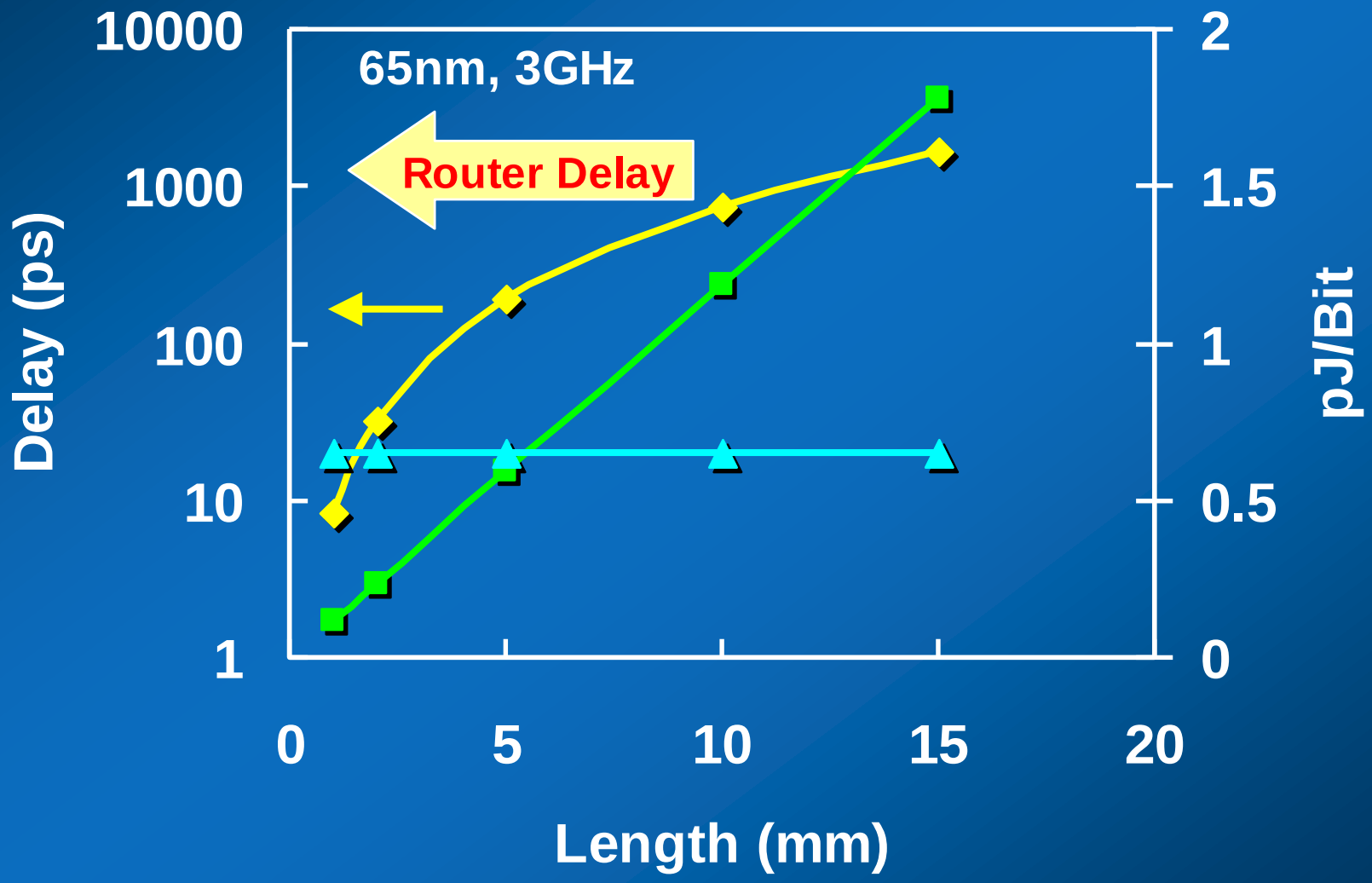
Point to point busses: fast signaling over longer distance

- Board level, between boards, and racks
- High frequency, narrow links
- 1D Ring, 2D Mesh and Torus to reduce latency
- Higher complexity and latency in each node

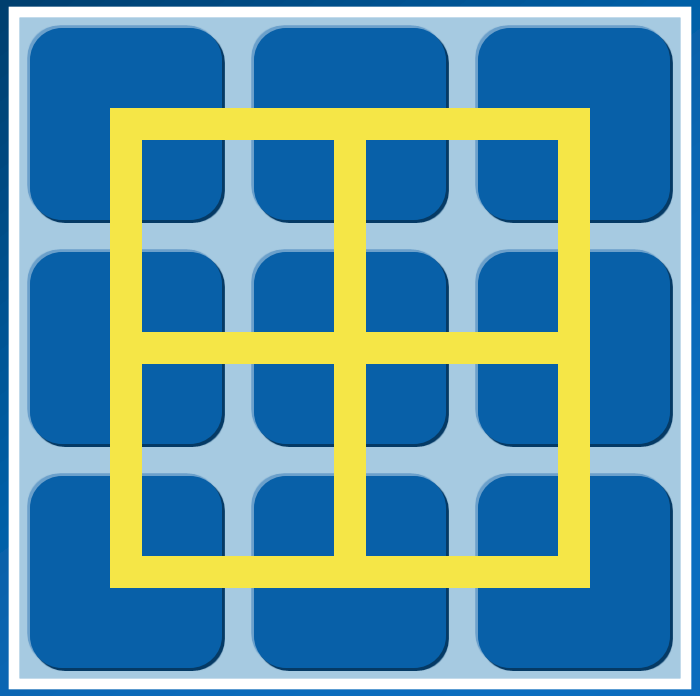
Hence, emergence of packet switched network

But, pt-to-pt packet switched network on a chip?

Interconnect Delay & Energy



Bus—The Other Extreme…



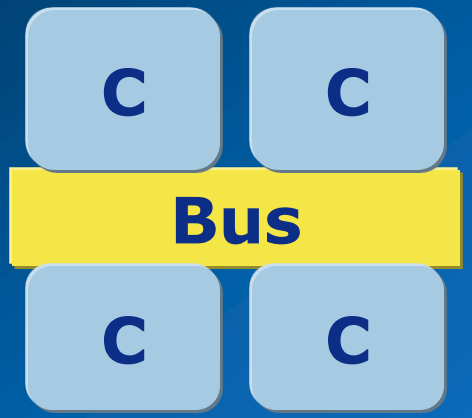
Issues:
Slow, < 300MHz
Shared, limited scalability?

Solutions:
Repeaters to increase freq
Wide busses for bandwidth
Multiple busses for scalability

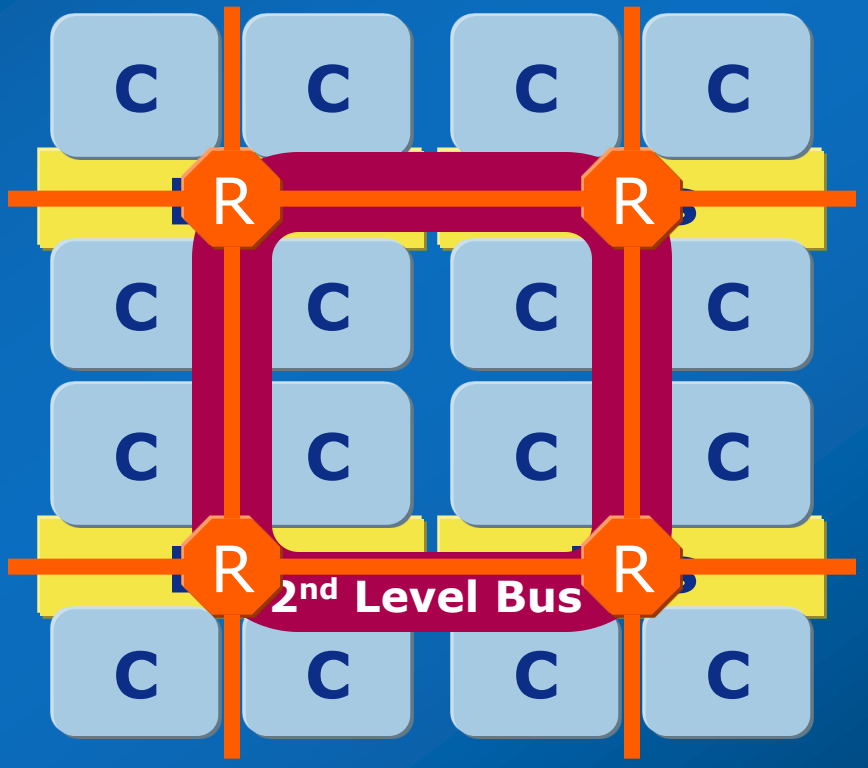
Benefits:
Power?
Simpler cache coherency

Move away from frequency, embrace parallelism

Hierarchical & Heterogeneous



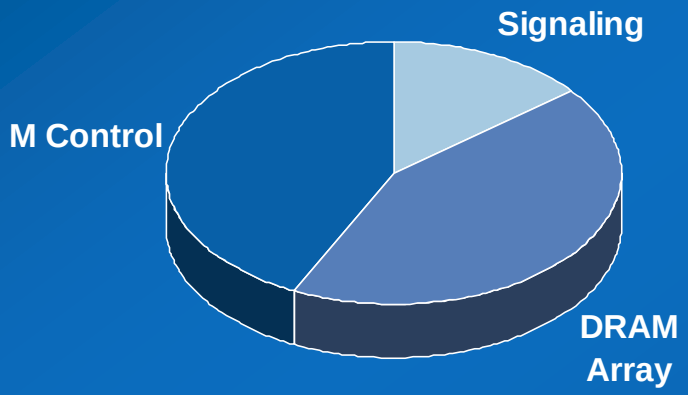
Bus to connect over short distances



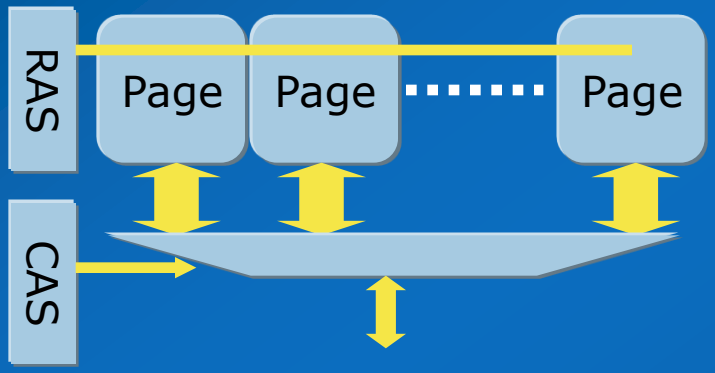
Hierarchy of Buses or hierarchical circuit and packet switched networks

Revise DRAM Architecture

Energy cost today:
~175 pJ/bit

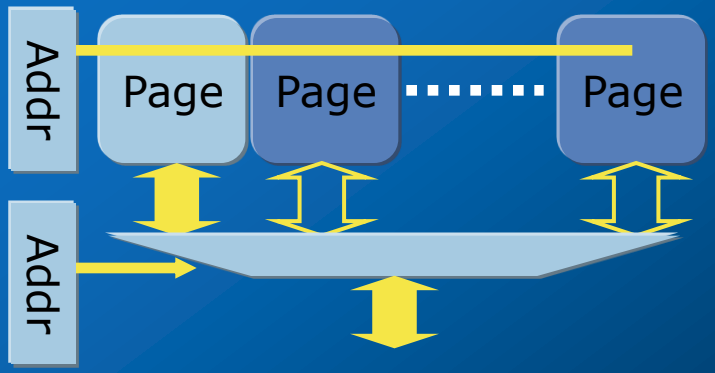


Traditional DRAM



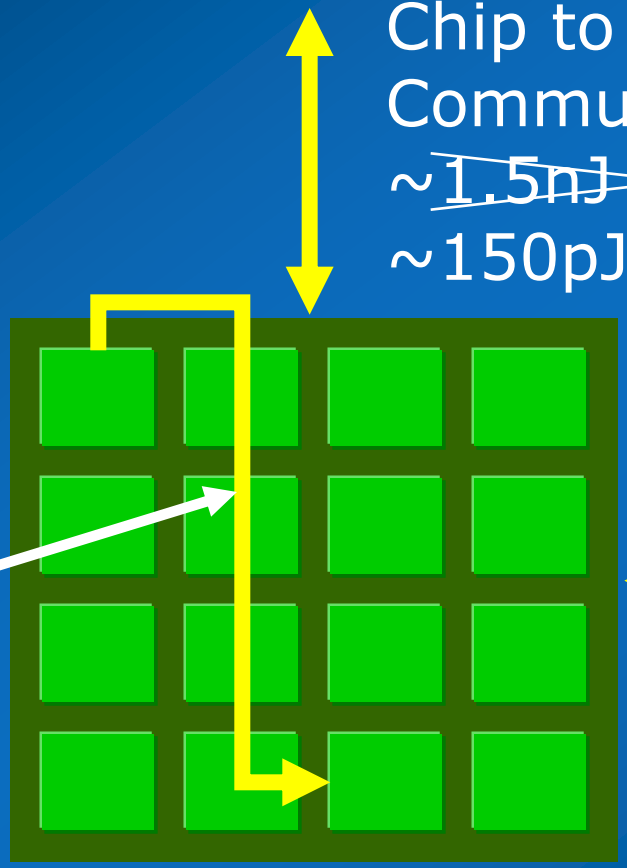
Activates many pages
Lots of reads and writes (refresh)
Small amount of read data is used
Requires small number of pins

New DRAM architecture



Activates few pages
Read and write (refresh) what is needed
All read data is used
Requires large number of IO's (3D)

Data Locality



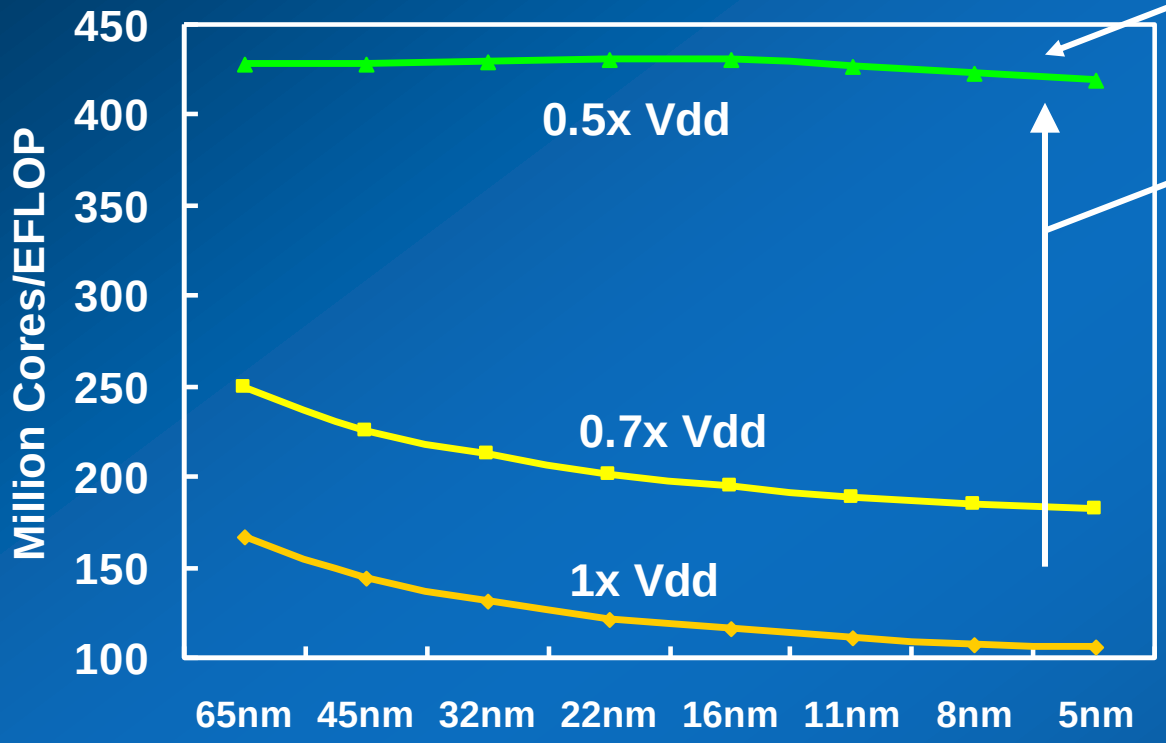
Chip to memory
Communication:
~~~1.5nJ per Byte~~~
~150pJ per Byte

Core-to-core
Communication
on the chip:
~10pJ per Byte

Chip to chip
Communication:
~100pJ per Byte

Data movement is expensive—keep it local
(1) Core to core, (2) Chip-to-chip, (3) Memory
~~~

Impact of Exploding Parallelism



Almost flat because Vdd close to Vt

4X increase in the number of cores (Parallelism)

Increased communication and related energy

Increased HW, and unreliability

1. Strike a balance between Com & Computation
2. Resiliency (Gradual, Intermittent, Permanent faults)

Road to Unreliability?

From Peta to Exa	Reliability Issues
1,000X parallelism	More hardware for something to go wrong >1,000X intermittent faults due to soft errors
Aggressive Vcc scaling to reduce power/energy	Gradual faults due to increased variations More susceptible to Vcc droops (noise) More susceptible to dynamic temp variations Exacerbates intermittent faults—soft errors
Deeply scaled technologies	Aging related faults Lack of burn-in? Variability increases dramatically

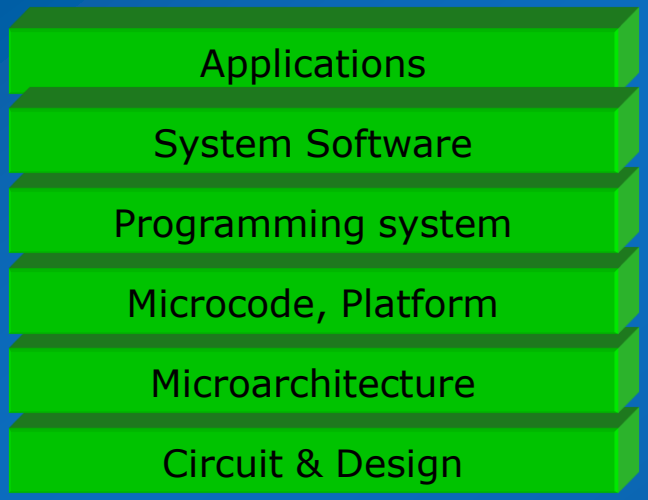
Resiliency will be the corner-stone

Resiliency

Faults	Example
Permanent faults	Stuck-at 0 & 1
Gradual faults	Variability Temperature
Intermittent faults	Soft errors Voltage droops
Aging faults	Degradation

Faults cause errors (data & control)	
Datapath errors	Detected by parity/ECC
Silent data corruption	Need HW hooks
Control errors	Control lost (Blue screen)

Minimal overhead for resiliency



Error detection
Fault isolation
Fault confinement
Reconfiguration
Recovery & Adapt

Needs a Paradigm Shift

Past and present priorities—

Single thread performance	Frequency
Programming productivity	Legacy, compatibility Architecture features for productivity
Constraints	(1) Cost (2) Reasonable Power/Energy

Future priorities—

Throughput performance	Parallelism
Power/Energy	Architecture features for energy Simplicity
Constraints	(1) Programming productivity (2) Cost

Evaluate each (old) architecture feature with new priorities

Summary

Von-Neumann computing & CMOS technology (nothing else in sight)

Voltage scaling to reduce power and energy

- Explodes parallelism
- Cost of communication vs computation—critical balance
- Resiliency to combat side-effects and unreliability

Programming system for extreme parallelism

System software to harmonize all of the above